Interchromosomal LD in HapMap Genotypes

Albert Vernon Smith

2007-02-19

1 Summary

Interchromosomal LD has been examined with HapMap genotypes. Starting with all genotypes in the non-redundant set of release 21, tag SNPs were chosen by pairwise tagging for all the autosomes. Using these tag SNPs, LD was measured for all interchromosomal tag SNP pairs for each of the three analysis panels. Interchromosomal SNP pairs with $r^2 \geq 0.8$ were examined in more detail.

2 Results

2.1 Tag SNPs

Starting with all the non-redundant genotypes from HapMap release 21, tags were chosen by pairwise tagging. SNPs were filtered to those with a MAF ≥ 0.05. Then, pairwise tags were chosen based on pairwise linkage disequilibrium (r^2) estimates for each analysis panel with an r^2 threshold of 0.8. This resulted in the number of tag SNPs show in Table 1.

panel	count
CEU	570,527
JPT+CHB	502,314
YRI	1,106,868

Table 1: Number of Tag SNPs for each of the three analysis panels.

2.2 Interchromsomal LD

Then, using all of these tags were used to calculate intrachromosomal LD within each panel. Analysis was done using an EM algorithm based on code from Goncalo Abecasis. Marker pairs with $r^2 \geq 0.6$ were stored for further analysis.

As shown in Table 2, the number of distinct pairs of SNPs with $r^2 \geq 0.6$ and ≥ 0.8 are shown. The numbers are shown for the original tag set. Additionally, to facilitate cross panel comparisons, all the SNPs which were captured by the original tag sets, were also analyzed for high interchromosomal LD. This extension was done because we wanted to investigate if any SNP pairs displayed high interchromosomal LD in multiple analysis panels. With initial analysis limited to tag SNPs, while specific regions might overlap, the exact SNPs used as tags might not be the same. The higher number of SNP pairs seen for the YRI panel, particularly at the lower threshold, presumably derives from the larger number of polymorphic SNPs in this panel, and is artifactual.

Since it was presumed that a large source of observed interchromosomal LD might come from SNPs which were mismapped, the maximum local r^2 value was determined. It was presumed that if (at least) one of a SNP pair was mismapped, then they would be much less likely to display high local LD for both SNPs.

panel	set	r^2 threshold $\geq$ 0.6	$\geq$ 0.8
CEU	tags	4,311	289
	all	22,963	2,751
YRI	tags	14,333	480
	all	42,302	2,127
JPT+CHB	tags	628	321
	all	3,552	2,090

Table 2: Number of SNP Pairs with high interchromosomal LD at different r^2 thresholds.

panel	interchrom $r^2 \geq 0.8$ all	local $r^2 \geq 0.6$
CEU	289	59
YRI	480	133
JPT+CHB	628	25

Table 3: Number of Tag SNP Pairs with high local LD.

As shown in Table 3, a large proportion of the SNPs with high interchromosomal LD, are filtered out as not having high local LD for both of the SNPs. This is suggestive that some of the interchromosomal LD observed derives from SNPs that are mismapped, or from assays which are reflecting other locations in the genome.

These results are visualized in Figures 1, 2, and 3.

2.3 Assay Alignment to Genome

panel	Primers Align	Primers Misalign
CEU	127	162
JPT+CHB	187	134
YRI	281	199

Table 4: Alignment of primers for Tag SNP pairs with $r^2 \geq 0.8$.

Since some of interchromosomal LD may come from assays where the primers are aligning to alternative locations in the genome, all primers from the relevant assays were aligned to the genome with BLAT. Then, the chromosomal locations for the best hits were compared to the location of the assay. As shown in Table 4, approximately 50% of the assays had one or more of the primers align to alternative locations in the genome. However, there in the cases of some misaligned primers, most the primers still aligned to the location surrounding the SNP, and only one primer from an assay was aligned to another location in the genome (data not shown). Additionally, in the cases of misalignment, the alternative location for the primers did not appear correlate with the location of the SNP showing high interchromosomal LD (data not shown).

2.4 Genome Duplications and Interchromosomal LD

Another possible explanation for why high interchromosomal LD is observered could be due to reasons relating to the underlying genomic structure. As annotated by Evan Eichler's group, the location of segmental duplications was compare to the location of the tags SNPs which show interchromosomal $r^2 \geq 0.8$. As

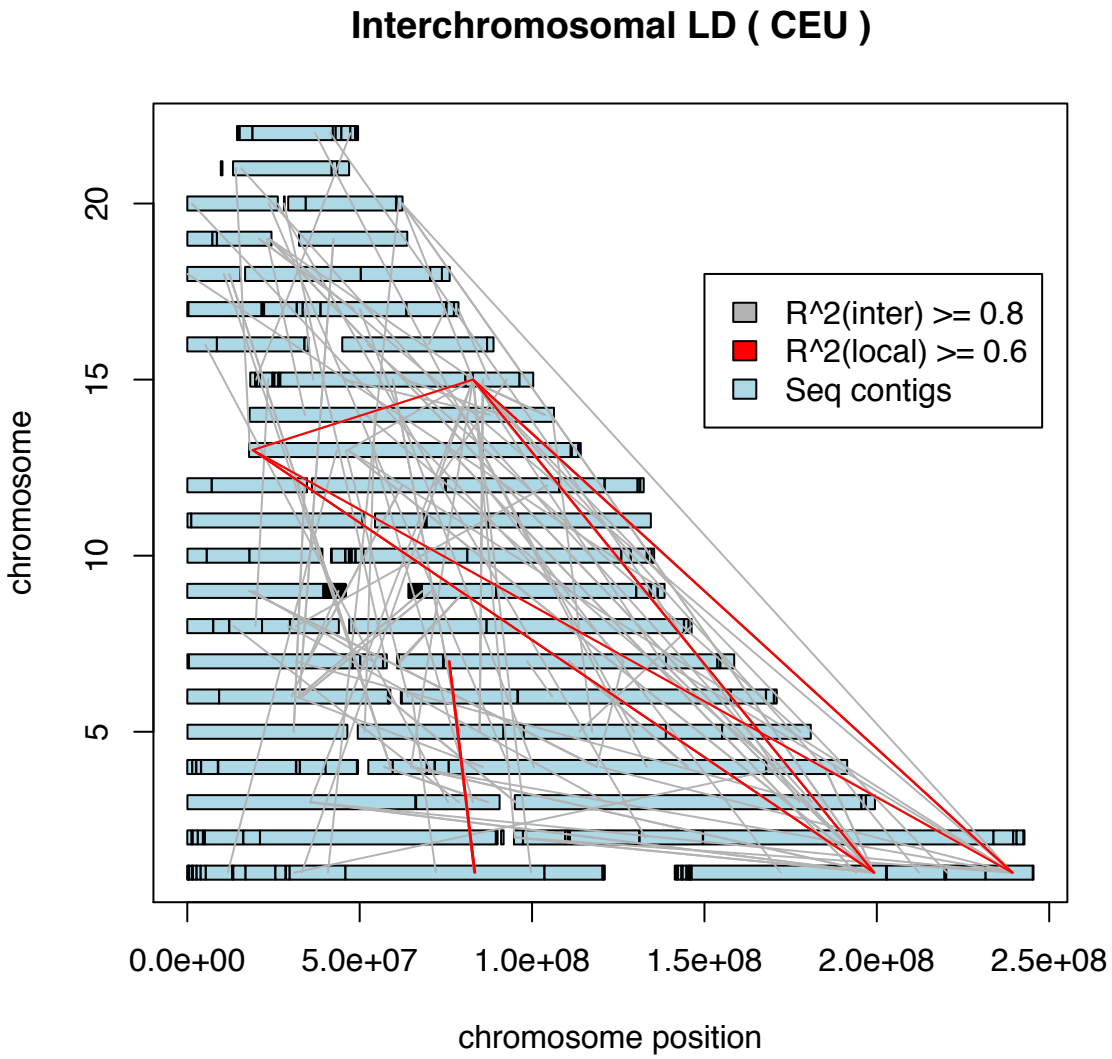

Figure 1: SNP Pairs with high interchromosomal LD in CEU panel

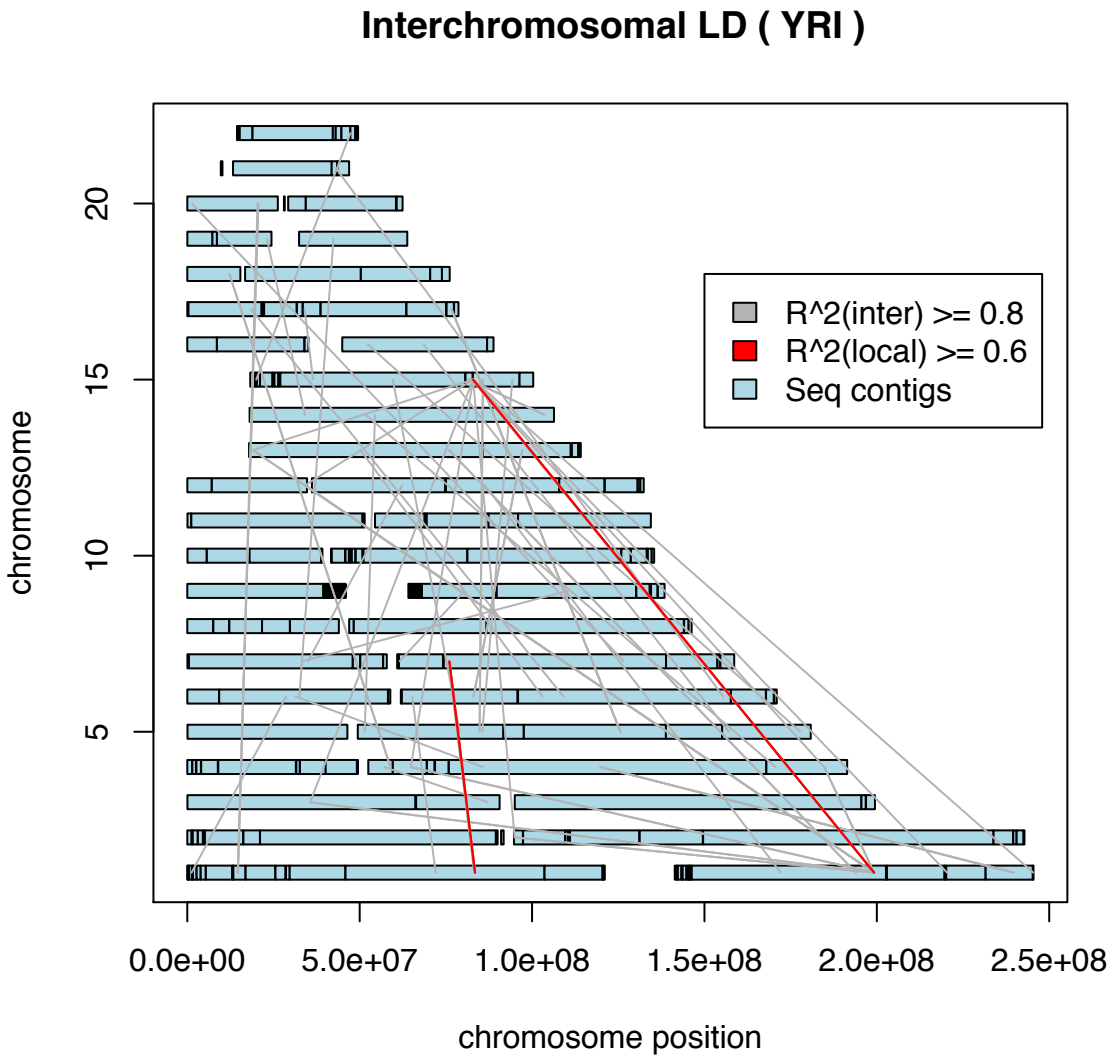

Figure 2: SNP Pairs with high interchromosomal LD in YRI panel

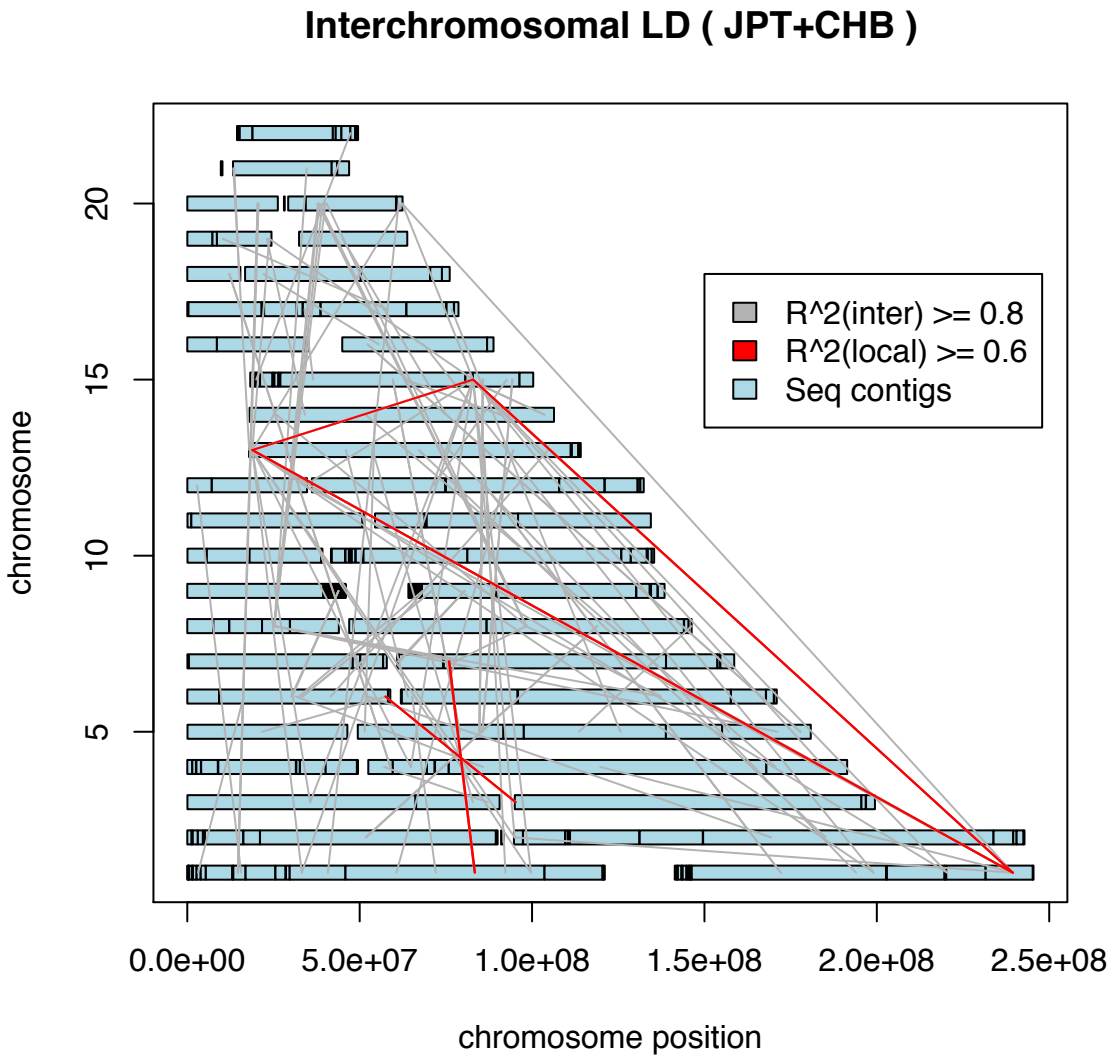

Figure 3: SNP Pairs with high interchromosomal LD in JPT+CHB panel

panel	Segmental Dup	HapMapable
CEU	108	181
JPT+CHB	84	237
YRI	72	408

Table 5: Overlap of Tag SNPs pairs with Segmental Duplications.

shown in table 5, approximately 25% of the observed SNP pairs are overlapping which annotated segmental duplications on one or both ends. This is a higher proportion than exists overall in the HapMap data, and suggests that a proportion of the observed interchromosomal LD is related to underlying duplications.

If a consistent underlying explanation for interchromosomal LD relating to genome structure and/or genome assembly issues, it would then be expected that SNP pairs related to those issues would more like appear in multiple of the analysis panels. Therefore, SNPs pairs (from the "all" set) were tested for those which are appearing repeatedly in multiple sets.

duplication status	all	local $r^2 \geq 0.6$
NONE	610	12
DUP (at least one)	366	83

Table 6: Overlap of SNPs pairs appearing in multiple populations with Segmental Duplications. SNP pairs are limited to those with interchromosomal $r^2 \geq 0.8$, and the local threshold is $r^2 \geq 0.6$.

These results are summarized in Table 6. SNPs were examined for SNP pairs which appears with $r^2 \geq 0.8$ in multiple panels. Pairs were also classified by whether they show high local r^2. Additionally, SNP pairs were classified for those which overlapped an annotated segmental duplication for one or both of the SNPs in the pair. As seen in Table 6, there are large proportion of the SNPs which are appearing in multiple populations, and these represent up to 50% of all SNP pairs which have interchromosomal $r^2 \geq 0.8$. The high overall proportion of SNPs pairs appearing in multiple populations suggests there is a common underlying cause. When looking at SNPs where one or both of the SNPs have low local LD, approximately 60% are not overlapping regions of none segmental duplications. This suggests the SNPs, sequence and/or assays should be assigned to different regions of the genome.

When limiting to those which also exhibit high local r^2, 87% align with annotated segmental duplications. This suggests that the reason underlying the observed interchromosomal LD in this category is largely related to the segmental duplications. These results are also presented visually in Figure 4. Of the 95 with high local r^2, several broad categories can be observed. There are 45 pairs which are between chr1 and chr7. Those two regions have been annotated as being segmental duplications with respect to one another. This suggests potential issues with assembly, or that these results arise from the duplication status. Additionally, there 20 SNP pairs which overlap annotated duplication (for at least one SNP) to chrY, particularly for the set which arises between chr1, chr13, and chr15. Those again suggest that these results are related to duplication status. The last clear category is a set of 21 SNP pairs between chr3 and chr6. For these, there is essentially no clear segmental duplication annotated for theses region of the genome by Evan Eichler. However, both regions are immediately sub-telomeric, and the chr6 region is contained within a known copy number variation. Again, this suggests the results are arising from the underlying genomic structure of the region.

3 Summary

Interchromosomal LD has been observed in the HapMap samples. For those SNPs which appear to be appriopriately map (based on high local r^2) and appearing in mulitple populations, the likely explanation

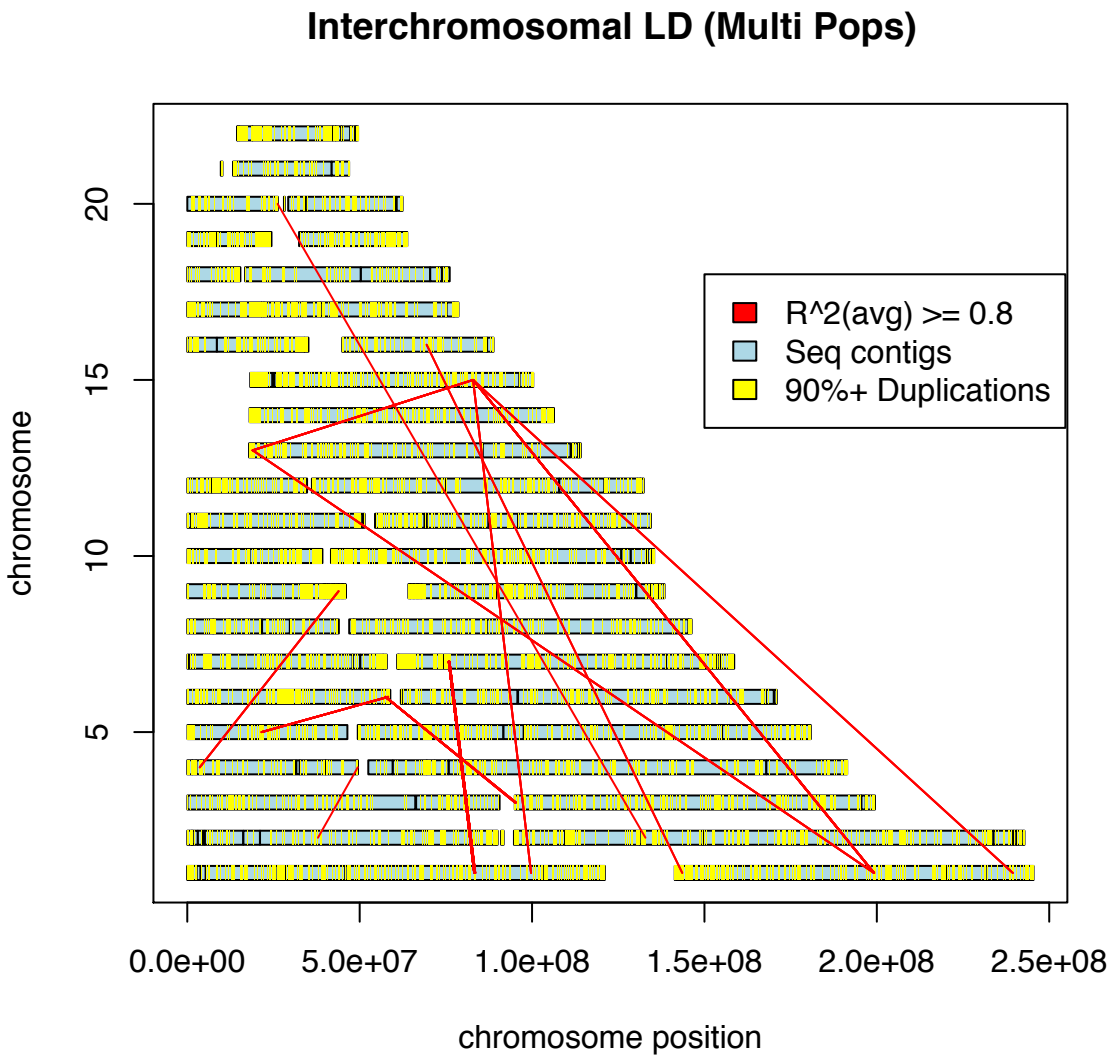

Figure 4: SNP Pairs with high interchromosomal LD in multiple analysis panels (Limited to pairs with local $r^2 \geq 0.6$ and MAF ≥ 0.10).

relates to the underlying genomic structure. There appears to be an additional set which likely results from mismapping of the SNPs and/or assays. While there are individual pairs beyond those which appear uniquely in each analysis panel, there is not a clear and consistent pattern suggestive of biological relevance beyond that arising from copy number variation and/or segmental duplication.